Energy-Efficient Cloud Radio Access Networks by Cloud Based Workload Consolidation for 5G

Tshiamo Sigwele[a,*], Atm S. Alam[a], Prashant Pillai[a], Yim F. Hu[a]

[a] *Faculty of Engineering and Informatics, University of Bradford, Bradford, BD7 1DP, West Yorkshire, United Kingdom*

Abstract

Next-generation cellular systems like fifth generation (5G) ~~is~~ *are* expected to experience tremendous traffic growth. To accommodate such traffic demand, there is a need to increase the network capacity that eventually requires the deployment of more base stations (BSs). Nevertheless, BSs are very expensive and consume a lot of energy. With growing complexity of signal processing, baseband units are now consuming a significant amount of energy. As a result, cloud radio access networks (C-RAN) have been proposed as an energy efficient (EE) architecture that leverages cloud computing technology where baseband processing is performed in the cloud. This paper proposes an energy reduction technique based on baseband workload consolidation using virtualized general purpose processors (GPPs) in the cloud. The rationale for the cloud based workload consolidation ~~technique~~ model is to switch off idle baseband units (BBUs) to reduce the overall network energy consumption. The power consumption model for C-RAN is also formulated with considering radio side, fronthaul and BS cloud power consumption. Simulation results demonstrate that the proposed scheme achieves an enhanced energy performance compared to the existing distributed long term evolution (LTE) RAN system. The proposed scheme saves up to 80% of energy during low traffic periods and 12% during peak traffic periods compared to baseline LTE system. Moreover, the proposed scheme saves 38% of energy compared to

☆This document is a collaborative effort.

*Corresponding author

Email addresses: T.Sigwele@bradford.ac.uk (Tshiamo Sigwele), A.S.Alam5@bradford.ac.uk (Atm S. Alam), P.Pillai@bradford.ac.uk (Prashant Pillai), Y.F.Hu@bradford.ac.uk (Yim F. Hu)

the baseline system on a daily average.

Keywords: Cloud Computing, C-RAN, Energy Efficiency, Workload Consolidation, Virtualization, 5G.

1. INTRODUCTION

Lately, the number of connected devices has grown into billions and today mobile operators are facing the serious challenge of ever increasing demand of high data rates. For example, Huawei Technologies envisages that 100 billion devices will be connected to the internet by 2020 [1]. This will cause a surge in global mobile voice and data traffic. This tremendous traffic growth is due to the introduction of smart phones and other high-end devices like the android, iphone, iPad, kindle and gaming consoles spawning a raft of data intensive applications, Internet of Things (IoT) and machine-to-machine connections. As a result, next-generation mobile communication networks such as fifth-generation (5G) have received exceptional expectations with targeting to increase 1000 fold capacity, 100 times data rate, and millisecond-level delay [2]. More base stations (BSs) with a mixer of macro and small cells are required to fulfil these increasing capacity demands. However, traditional BSs consume a significant portion of energy in cellular network, estimated around 60-80% of the whole network energy consumption [3]. In addition, the BSs are also expensive as well as energy inefficient due to their operational design and dynamic nature of traffic demand in both temporal and spatial domains called tidal effect as shown in Fig. 1 [4]. These BSs have been preconfigured to provide peak capacities and their baseband processing capacity is only being used for its own coverage rather than being shared in a large geographical area. The baseband processors are always on irrespective of traffic needs causing low utilization, waste of energy and processing resources [5]. Not only that, BSs also causes a greater impact to the environment by emitting large amounts of CO_2 and contributes to the mobile networks operating expenditure (OPEX). Therefore, it is important to solve this problems by taking advantage of spatial and temporal dynamic nature of traffic to develop energy efficient mechanisms in BSs that scale with traffic demand during low traffic periods.

Within each BS, a large amount of power is consumed by the power amplifier (PA) and the baseband unit (BBU) or computing servers as shown in Fig. 2. The energy consumption of BBUs is getting more and more dominant

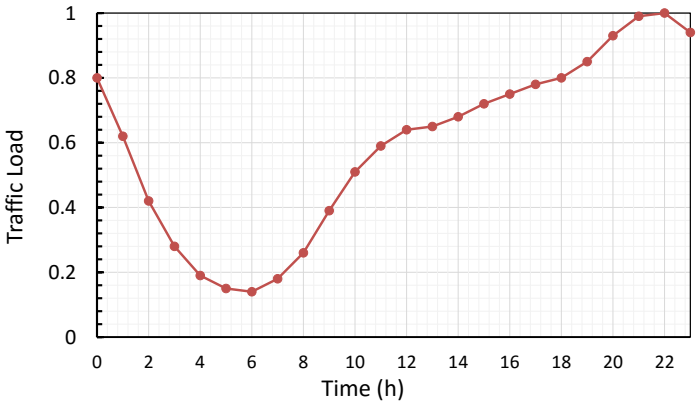

Figure 1: Typical residential BS traffic profile [4]

in small cells due to gradual shrinking of cell size and the growing complexity of signal processing [3][6]. Hence, it is crucial to optimize energy efficiency (EE) in the BBU servers which is the target for this paper. Other BS power consuming components include main supply, direct current converter (DC-DC), radio frequency (RF) module and cooling only for macro base stations. Many energy-efficient schemes for wireless systems have been implemented such as BS sleeping where offloading traffic to neighbouring BSs and then completely turning off the BS during low traffic is implemented [7], discontinuous transmission (DTX) where a BS is temporally switched off without offloading [8], cell zooming where cell size changes [9], and utilizing renewable energy sources [10].

Cloud radio access networks (C-RAN) have been recently proposed as a promising solution for reducing energy usage within the cellular networks by leveraging cloud computing technology [3]. This paper extends our previous research in [4] and [11] where approximation heuristic bin-packing algorithms were proposed with the aim of minimizing energy consumption in C-RAN by reducing the number of cloud BBUs used. The main contribution in this paper are as follows:

(i) It is not practical to use the previous ~~simplitic~~ *simplistic* power consumption models for C-RAN. A new power consumption model for C-RAN comprising of the separation of radio remote radio head (RRH) and baseband processing units is derived. Previous BS power models can not be used because in C-RAN baseband processing power, cooling and housing are shared in the cloud.

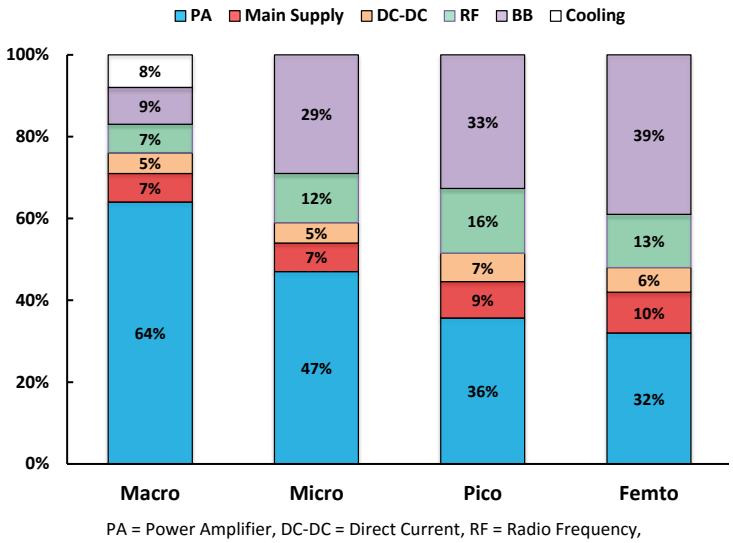

Figure 2: BS Power Consumption for different cell sizes [3]

(ii) An energy-efficient algorithm through cloud based workload consolidation ~~technique~~ model is proposed for reducing energy consumption on the cloud part of the C-RAN architecture. In this scheme, traffic workloads are distributed among the BBU cloud servers such that each server operates at its full utilization. As such, idle BBU servers are turned off while utilization is maximized hence improving the overall network EE.

(iii) The simulations results validates the energy-efficiency improvement of C-RAN using the proposed workload consolidation ~~technique~~ *framework* which is then compared with traditional long term evolution (LTE) system comprising of individual BBU servers within each cell.

The rest of the paper is organised as follows: Section II discusses the related works while the baseband workload consolidation framework for C-RAN is described in Section III. Section IV provides the simulation results and discussion, while Section V provides some concluding comments.

2. RELATED WORK

There have been some previous works on energy-aware algorithms in C-RAN. Authors in [12] proposed a BBU reduction scheme for C-RAN that dynamically allocates BBUs to RRHs based on the imbalance of subscribers

in office/residential areas. A set of upper limit of BBU utilization is defined to avoid overloading of the BBU. Even though the reduction of the number of BBUs required is achieved, the model performs poorly during high traffic loads. This is due to the high consumption of power since more BBUs are allocated to meet traffic demands at high traffic load. Authors in [13] proposed a C-RAN system using virtualization technology on general purpose processors BBUs are dynamically provisioned according to traffic load. However, the paper fails to show how the number of BBUs are reduced with dynamic traffic load. Also, ~~linux~~ *Linux* operating system assisted virtualization is used which add more delays and jitter due to virtualization platform when performing baseband processing on virtual BSs. In [14], the authors proposed an analytical energy model of a computational-resource-aware virtual BS in a cloud-based cellular network architecture. The authors consider the energy-delay trade-offs of a virtual BS considering BS sleeping mode in general purpose platforms. The paper does not show how the energy savings of the virtual BS model scales with traffic load. In [15], L. Cheng *et al.* developed an energy efficient C-RAN ~~systemd~~ *system* with a reconfigurable backhaul that allows 4 BBUs to connect flexibly with 4 RHHs using radio-over-fiber technology. The backhaul architecture allows the mapping between BBUs and RRHs to be flexible and changed dynamically to reduce energy consumption in the BBU pool. However, the paper assumes static user traffic whereas in reality BS traffic is dynamic. S. Namba *et al.* in [16] ~~proposes~~ *proposed* a BBU reduction network architecture called Colony-RAN due to its ability to flexibly change cell layout by changing the connections of BBUs and RRHs in respect to traffic demand. However, the proposed method has frequent ping pong reselections of RRH to BBU.

Liming Cheng et al. in [17] focused on the spectral efficiency (SE) and EE of C-RAN implementation. The SE advantage was achieved by cooperative transmission among RRHs while EE performance was improved through proposed computational efficient pre-coding scheme. Nonetheless, the paper assumes a realization of C-RAN in which all information is available which might bring about high bandwidth overheads in the fronthaul. The author in [18] investigates the cooperative transmission design for C-RAN considering fronthaul capacity and cloud processing constraints. The author considers the joint transmission scheme where the baseband signals and precoding vectors are processed and calculated by the cloud.

The author in [19] reduces the network cost and energy consumption in C-RAN by dynamically allocating centralized BBU resources to RRHs de-

pending on the traffic conditions, however, the power consumption on RRHs and BBUs are assumed to be static and are independat of traffic load which is not realistic. Authors in [20] proposed a C-RAN system using virtualization technology on general purpose processors where BBUs are dynamically provisioned according to traffic load. However, the paper fails to show how the number of BBUs are reduced with dynamic traffic load. In [21], L. Cheng et al. developed an energy efficient C-RAN system with a reconfigurable backhaul that allows 4 BBUs to connect flexibly with 4 RHHs using radio-over-fiber technology. The backhaul architecture allows the mapping between BBUs and RRHs to be flexible and changed dynamically to reduce energy consumption in the BBU pool. However, the paper assumes static user traffic whereas in reality BS traffic is dynamic.

The author in [22] proposes an energy efficient scheme with the regularized zero-forcing precoding for the distributed large-scale multiple input multiple output (MIMO) C-RAN which consists of a large number of spatially distributed remote radio heads (RRHs) and the simulation results show that the proposed scheme in [22] provides better EE with the consideration of QoS support. Y. Ma et al. [23] developed a new precoding scheme based on the Lagrangian dual relaxation and simulation results proved that the energy performance is better than the conventional one.

3. BASEBAND WORKLOAD CONSOLIDATION FRAMEWORK FOR C-RAN

3.1. Basics of C-RAN Architecture

The C-RAN architecture adopted in this paper is shown in Fig. 3. C-RAN comprises of the 4 Cs which stand for centralized, collaborative, cooperative and clean/green [3][5]. The BBUs are separated from the cell areas and centralized in the BS cloud ~~infranstracture~~ *infrastructure* leaving only the less intelligent RRH in the cell sites. Digital baseband processing is ~~perfomed~~ *performed* in the cloud on virtualised servers while RRH perform radio frequency functions, analogue to digital conversions and vice-versa and up-down conversions. The RRH and BS cloud are connected by high bandwidth fiber fronthaul. The main advantages of C-RAN architecture are [3]: (i) reduction in air conditioning and other ~~onsite~~ *on-site* power-consuming equipments, (ii) ease of future extension of the network simply by installing new RRHs and connecting them to the BBU pool to expand the network coverage or ~~spliting~~

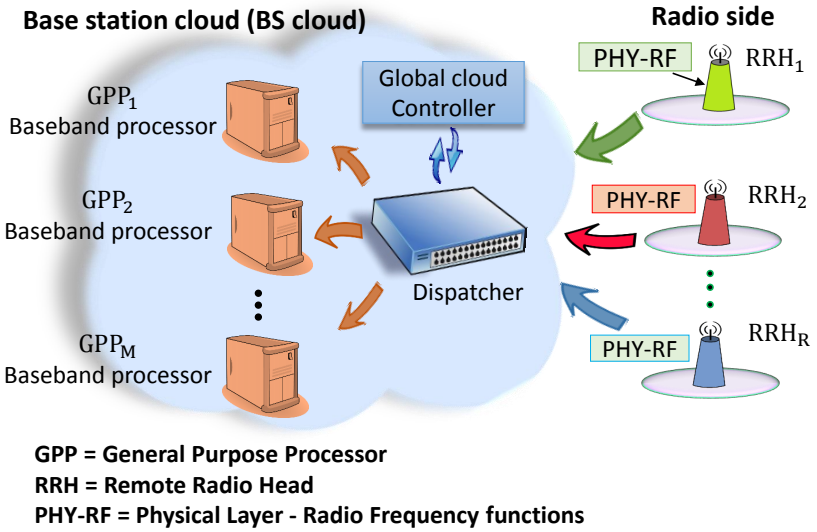

Figure 3: An illustration of a C-RAN architecture.

splitting the cell to improve capacity, (iii) enabling coordinated multipoint (CoMP) and easy ~~intercell~~ *inter-cell* interference coordination (ICIC), (iv) inter-operator sharing of ~~infranstracture~~ *infrastructure* i.e. BBUs and RRH sharing, and (vi) vendor-agnostic (interoperable), open and programmable. The main drawback for C-RAN is that the fronthaul link has high bandwidth and low latency requirements. For example, in the extreme case, a time division LTE 8 antenna with 20MHz bandwidth will need a 10 Gbps transmission rate.

3.2. System Model

Consider an LTE C-RAN downlink system consisting of a set of RRHs $\boldsymbol{\mathcal{R}} = [RRH_n : n = 1, 2, ..., R]$, where R is the maximum number of RRHs each serving a cell. Define a set of users in the entire network as $\boldsymbol{\mathcal{U}}$. It is also assumed that BBU processors used in C-RAN are the general purpose processors (GPPs) ~~centraly~~ *centrally* located in the BS cloud. The GPPs are used for baseband signal processing in the cloud due to their affordability, programmability and high processing capabilities. The costly and non ~~programable~~ *programmable* traditional digital signal processors (DSP) are replaced by GPPs. The BS cloud comprise of many GPPs with the ability to process any signal from the radio side. It is assumed that the GPPs in the

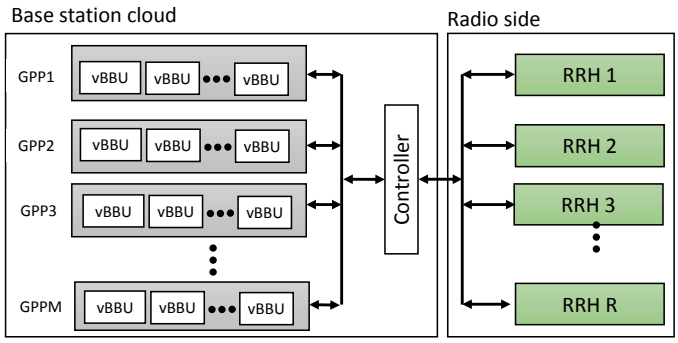

Figure 4: *System Model.*

pool are denoted by a set, $\boldsymbol{\mathcal{M}} = [GPP_i : i = 1, 2, ..., M]$, where M is the total number of GPPs for processing baseband signals of R cells and M is to be minimised. The workload from RRHs is routed via a dispatcher which distributes the workload among GPPs. The global cloud controller (GCC) manages all control operation in the cloud and keeps track of traffic load in the network and it is where the workload consolidation algorithm is located which will be described in section III. The power consumption of the dispatcher and the global controller ~~are~~ *is* assumed to be negligible. *Also consider that the virtual machines (VMs) running on the GPPs are denoted as virtual BBUs(vBBUs) and each RRH has its own specific vBBU. Therefore assume a set of vBBUs as* $\boldsymbol{\mathcal{V}} = [vBBU_j : j = 1, 2, ..., R]$ *where* R *is the total number of vBBUs in the BS cloud. Each vBBU is atomic and can be processed by only one GPP. The system model is shown in Figure 4.*

3.3. Proposed Power Consumption Model

First, the power model for traditional LTE system is formulated as baseline. The model is derived from the EARTH project [24] and it was found that the power consumption of an eNodeB P_{eNodeB} can be approximated as:

$$P_{eNodeB} = \begin{cases} P_0 + \triangle_P . P_{max} . \rho_n; & \text{if } \ 0 < \rho_n \leq 1 \\ P_{sleep}; & \text{if } \ \rho_n = 0 \end{cases} \tag{1}$$

where P_0 is the static load ~~independant~~ *independent* share from main power supply, baseband processing and cooling. The term $\triangle_P$ denote a power gradient variable of a particular BS, P_{max} denote the maximum transmission power when cell load is 100%. P_{sleep} is power consumption when the BS is in

sleep mode with 0% traffic load. The scaling parameter ρ_n is the normalized cell traffic load of the n^{th} BS, where $\rho_n = 1$ indicates a fully loaded system, e.g. transmitting at full power and full bandwidth, and $\rho_n = 0$ indicates an idle system. Thus, the total power consumption of the entire network for the baseline traditional LTE system $P_{baseline}$ is then formulated as:

$$P_{baseline} = \sum_{n=1}^{R} (P_0 + \triangle_P . P_{max} . \rho_n) \, ; \quad 0 < \rho_n \leq 1 \tag{2}$$

The LTE baseline model cannot be directly used in C-RAN because of the centralised BBUs and the shared housing, cooling and baseband processing power. As such, a new power model for C-RAN *will be formulated and derived in this paper* . The *proposed* power consumption model for C-RAN is divided into three separate parts: (i) radio side power consumption (P_{radio}) which is a sum of RRH power consumption, (ii) fronthaul power consumption ($P_{fronthaul}$) and (iii) BS cloud power consumption ($P_{BScloud}$).

$$P_{C-RAN} = P_{radio} + P_{fronthaul} + P_{BScloud} \tag{3}$$

(i) Radio side power consumption: The power consumption for the radio part can be calculated as:

$$P_{radio} = \sum_{n=1}^{R} (P_{static} + \triangle_P^{RRH} . P_{max}^{RRH} . \rho_n^{RRH}) \tag{4}$$

where P_{static} is load ~~independant~~ *independent* power consumption. There is no cooling losses on the RRH as cooling is done by natural air. The right term in (4) is the power consumption that depend on the traffic load of the RRH. The terms $\triangle_P^{RRH}, P_{max}^{RRH}$ and ρ_n^{RRH} are as described as in (1) but for RRH.

(ii) Fronthaul power consumption: The fronthaul is assumed to be fiber connection. Each RRH is connected to the BS cloud by a single mode 1310nm fiber. Power budget for the fiber is adopted based on [25]. It is assumed that the maximum RRH-BS cloud distance is 20km with 4 connector pairs per RRH-BS cloud connection, connector loss of 0.75dB per connector and 4 splicess with a loss of 0.25dB per splice. Using this values, the power budget for a single 20km fiber connection is approximately 5.39dBW.

(iii) BS cloud power consumption: The total power consumption in the BS cloud comprise of cooling power $P_{cooling}$ as well as the sum ~~off~~ *of* all

active GPPs power consumptions as follows:

$$P_{BScloud} = P_{cooling} + \sum_{i=1}^{M}(P_{GPP_i}) + P_{dispatcher} \tag{5}$$

where $P_{dispatcher}$ *is the power consumption of the dispatcher switch and* P_{GPP_i} is power consumption of the i^{th} active GPP. *The power consumption of the dispatcher is [31]:*

$$P_{dispatcher} = P_{base} + P_{config} \tag{6}$$

where P_{base}*,* P_{config} *and* $P_{control}$ *are the base power, configuration power (number of active ports) and power consumption of the control traffic.* The power consumption model of a standard GPP is as follows [26]:

$$P_{GPP_i} = P_0^{GPP_i} + \triangle_P^{GPP_i}.P_{max}^{GPP_i}.\rho_{GPP_i} \tag{7}$$

where $P_0^{GPP_i}$, $P_{max}^{GPP_i}$, and $\triangle_p^{GPP_i}$ denote idle mode power consumption, i.e., 0% central processing unit (CPU) utilization, the maximum power consumption of the GPP at 100% CPU utilization and the GPP power gradient, which is dependent on the type of GPP, respectively. The parameter ρ_{GPP_i} denotes the CPU utilization of GPP_i.

3.4. BBU Workload Consolidation Technique

The workload consolidation model is shown in Fig. 4. Workload consolidation is a cloud computing technique for processing workload into fewer number of computing servers to save energy by switching off underutilized servers. Workload in the context of C-RAN means baseband CPU processing power and the workload is measured in giga operations per second (GOPS). In simple terms, workload is the amount of CPU power required for processing cell traffic. In the BS cloud, a single GPP can be shared by many RRH coverage areas due to the programmable ~~capabily~~ *capability* of the GPPs. *As seen in Figure 5, traffic requests from the users in the coverage area are forwarded to the RRHs via the air interface. The RRHs then forwards the the user traffic to the BS cloud via the high bandwidth low latency fronthaul. Before a request is processed in the BS cloud, the request's control signals are forwarded to the GCC controller where admission control is performed to check whether there are enough resources in the BS cloud to process the request. When there are enough resources, the request is accepted and the data*

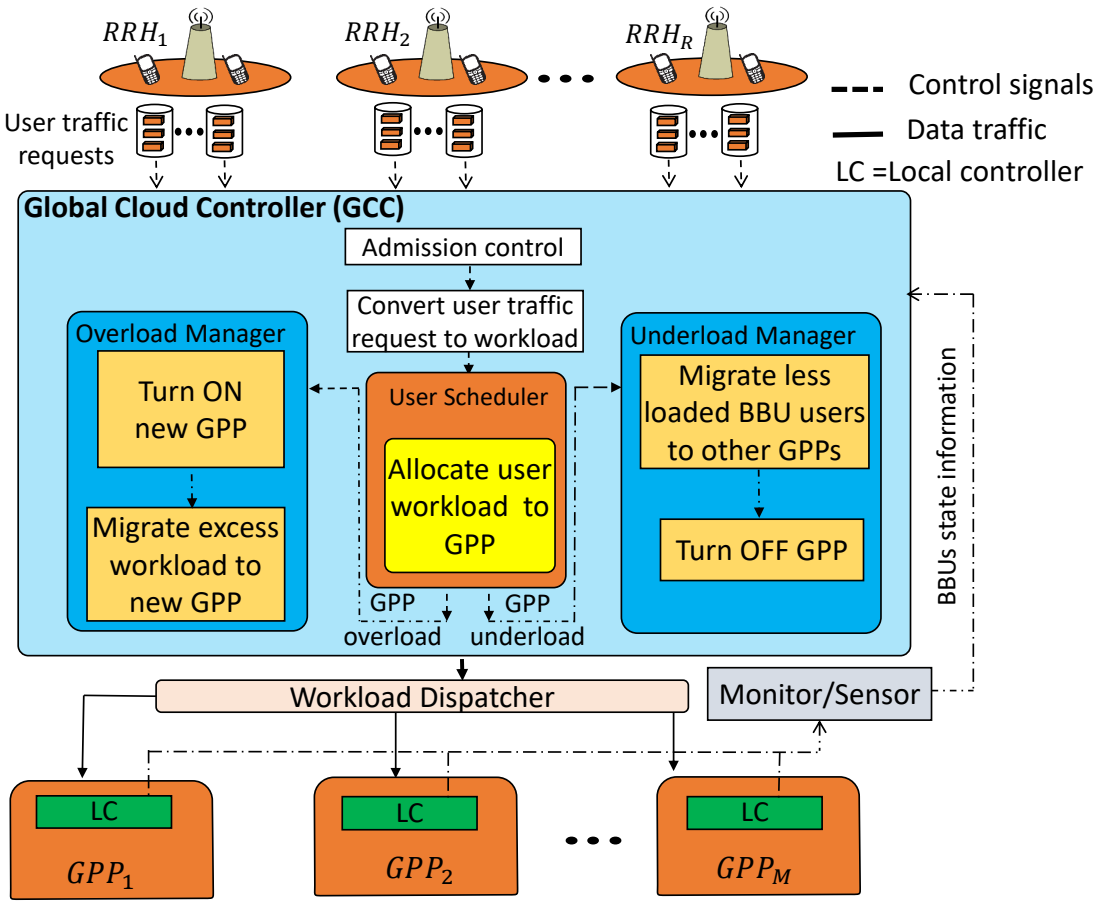

Figure 5: The proposed C-RAN workload consolidation model.

is forwarded via the dispatcher to a respective GPP that is not overloaded, otherwise the request is dropped. Data signals do not pass through the GCC controller, only control signals pass through the GCC controller as such, when the GCC controller is down, the dispatcher can still forward data to the GPPs even though data wont be allocated effectively to GPPs compared when there is the GCC. The aim is to pack BS traffic load into fewer number of GPPs and the problem is formulated as a bin-packing problem, which have been proved to be NP-hard. In our previous works [5][11], approximation heuristic bin packing algorithms like next fit, first fit, first fit decreasing have been proposed to minimise the number of GPPs in the BS cloud to save energy.

In this paper, the number of GPPs will be minimised by distribution of

user workload to GPP servers using the full bin packing allocation (FBPA) where user workloads are allocated to GPPs such that the GPPs are always fully utilised as explained later in this paper. With such a technique, less GPPs will be used and the remaining GPPs can be switched off. Different modules of the model described as follows:

(1) RRH: The RRH covers a cell where users are located. Users generate traffic requests, via the air interface to the RRH which are then forwarded to the GCC. The user traffic requests states some QoS related information like the number of physical resource blocks (PRBs), modulation and coding rate in the downlink and so on.

(2) Global cloud controller (GCC): The GCC is a centralised controller and is the main module located in the BS cloud. The GCC receives BBUs state information from GPPs containing their CPU utilization status and also user receives user requests from RRHs and make decisions as described below.

- **User traffic request to CPU workload converter:** The user request is converted to CPU workload W_u in GOPS because users are scheduled to GPPs based on CPU resources. In the later section, the corresponding ~~formular~~ *formula* will be provided.

- **User scheduler:** The user workload in GOPS is then allocated to a GPPs using FBPA (algorithm 2) such that maximum utilization in all GPPs is maintained.

- **Overload manager:** Overloading is GPP utilization is above maximum threshold. Overloading of GPPs normally happens when traffic load profile increases requiring more GPPs to be turned on. During GPP overloading, some excess users workload are migrated to new GPPs.

- **Underload manager:**Underload ~~occures~~ *occurs* when CPU utilization is below the minimum threshold. If traffic load in the coverage area is low, the processing workload will also be reduced. In such conditions, the underutilized (below threshold) GPPs are then turned off to save energy by issuing a ~~shutdown~~ *shut-down* GPP command.

(3) Workload dispatcher: The dispatcher receives requests control from the GCC on where to route the data from RRH to GPP. User data is then directed from RRH to a specific GPP without passing through GCC.

(4) Local controller (LC) and Monitor/Sensor: The LC is located in each GPP and regularly collects utilization status from the GPPs and forwards to the monitor/sensor module which then send status feedback to the GCC to check for overload or underload conditions of GPPs.

3.4.1. Mathematical Formulation

The user traffic dynamics from RRHs need to be converted to CPU processing resources i.e., the workload W_u in GOPS. A model for converting user traffic dynamics from cell areas to baseband CPU processing power has been proposed in [27] which has been adapted in this paper. This model states that the computing power in GOPS for a single user $u \in \boldsymbol{\mathcal{U}}$ at time t is calculated as:

$$W_{u,t} = \left(30Ant + 10Ant^2 + \frac{20KDL}{6}\right).\frac{R_{u,t}}{50} \tag{8}$$

$$\text{and,} \qquad W_n = \sum_{u \in \boldsymbol{\mathcal{U}}} W_{u,t} \tag{9}$$

where Ant is the number of antennas used per user, K is the modulation bits, D is the coding rate used, L is the number of multiple input multiple output (MIMO) layers used, $R_{u,t}$ is the number of PRBs and B is the bandwidth. The variable W_n denote the total workload for RRH_n. The total workload W_{total} from all cells in the entire network is then formulated as:

$$W_{total} = \sum_{n=1}^{R} W_n \tag{10}$$

It is important to note that this workload should not exceed the total BS cloud capacity of $(M * C_{cap})$ where C_{cap} is the maximum capacity of a single GPP. The minimum number of GPPs (operating at maximum utilization) M_{min} required to process the total workload W_{total} is approximated as:

$$M_{min} = \lceil \frac{W_{total}}{C_{cap}} \rceil \tag{11}$$

where the function $\lceil \cdot \rceil$ denotes the floor function for rounding up the value to the nearest upper integer. *This formula calculates the minimum number of GPPs that are turned on in the BS cloud when GPPs are operating at maximum CPU capacity. There will be an extra GPP which does not operate*

at full utilization and will give the inner part of the equation a decimal, as such the floor function will count the extra GPP such that the value of the formula is not a decimal but an integer of GPPs. It is assumed that all GPPs have the same capacity C_{cap} and consume the same energy for a given load. Thus, the CPU utilization ρ_{GPP_i} of each GPP can be written as:

$$\rho_{GPP_i} = \frac{W_{total}}{C_{cap}.M_{min}} \text{x} 100\% \tag{12}$$

3.4.2. C-RAN Workload Consolidation Algorithm

The proposed workload consolidation algorithm is shown in Algorithm 1. The algorithm takes the user $u \in \boldsymbol{\mathcal{U}}$ traffic request (number of PRBs required, the modulation bits, and the coding rate, MIMO layers, number of antennae) from a set of RRHs $\boldsymbol{\mathcal{R}}$ as inputs. The output is the minimum number of active GPPs M_{min}. For simplicity, the time parameter (t) is omitted. When the system is running well with no GPP workload overloading or underloading, then there are no users to migrate in between GPPs (line 1). A new user traffic request is converted to CPU baseband processing power W_u in GOPS using (7) and allocated to GPPs using FBPA in Algorithm 2 to be explained later in this section (lines 2-5). All GPPs are monitored and checked for overloading and underloading of workload (line 6). If a GPP is overloaded, the excess workload above the threshold are allocated to other GPPs (lines 7-9) using Algorithm 2. Invoking Algorithm 2 might end up turning a new GPP ON if all GPPs can not accommodate the excess users to be migrated (line 10). In such a case, the number of GPPs are incremented (line 11) and the variable $usersToMigrate$ is set to zero after all excess users are migrated (line 13).If a GPP is underloaded below a threshold (line 15), all users ~~workoad~~ *workload* in that GPP are stored on the variable $usersToMigrate$ (line 16), and allocated to highly loaded GPPs using Algorithm 2 (line 17) such that GPPs are always highly utilized. As such, the GPP becomes idle and it is turned off to save energy (line 18) resulting in decrementation of the active GPPs (line 19). Finally the variable $usersToMigrate$ is set to zero.

Algorithm 2 is the allocation algorithm of users workload W_u to GPPs in both underloaded and overloaded conditions. It takes users workload W_u and set of GPPs $\boldsymbol{\mathcal{M}}$ as inputs and the GPP allocation map as the output. The GPP allocation map is an association between user and GPP stating which user is allocated to which GPP. It is a matrix of size $|\boldsymbol{\mathcal{U}}|\text{x}M$ and each element of the matrix is either 1 where GPP_i processes workload W_u or 0

Algorithm 1 C-RAN Workload Consolidation Algorithm

Input: user $u \in \boldsymbol{\mathcal{U}}$ traffic request, GPP list $\boldsymbol{\mathcal{M}}$
Output: Minimum Active GPPs M_{min}

1: $usersToMigrate$ = NULL
2: **for** each user u in $\boldsymbol{\mathcal{U}}$ **do**
3: Convert user traffic request to CPU workload W_u
4: Allocate W_u to GPPs using ***Algorithm 2***
5: **end for**
6: **for** each GPP_i in $\boldsymbol{\mathcal{M}}$ **do**
7: **if** GPP_i is overloaded **then**
8: $usersToMigrate$ = get excess users from GPP_i
9: Allocate this users to GPPs using ***Algorithm 2***
10: **if** new GPP is activated using ***Algorithm 2*** **then**
11: $M_{min} = M_{min} + 1$
12: **end if**
13: $usersToMigrate.clear()$
14: **end if**
15: **if** GPP_i is underloaded **then**
16: usersToMigrate = get all users from GPP_i
17: Allocate this users using ***Algorithm 2***
18: Turn off this GPP
19: $M_{min} = M_{min} - 1$
20: $usersToMigrate.clear()$
21: **end if**
22: **end for**

otherwise. The notation $|\boldsymbol{\mathcal{U}}|$ represents the cadinality of $\boldsymbol{\mathcal{U}}$, which is the total number of users. In Algorithm 2, the GPPs in $\boldsymbol{\mathcal{M}}$ are sorted in decreasing order of CPU usage and stored in set $\boldsymbol{\mathcal{S}}$ (line 1). Then for each user workload W_u to be scheduled, it is scheduled on the left most GPP in set $\boldsymbol{\mathcal{S}}$ that has enough baseband processing resouces (lines 3-5), else a new GPP is activated (line 7) if all other GPPs are fully loaded usually during high traffic periods. The user workload is then scheduled on that new GPP (line 8).

3.5. ~~Perfomance~~ Performance Metrics

The following ~~perfomance~~ *performance* metrics will be used for evaluating the ~~perfomance~~ *performance* of the proposed LTE C-RAN workload

Algorithm 2 Full Bin Packing Algorithm (FBPA)

Input: user workload W_u, GPP list $\boldsymbol{\mathcal{M}}$
Output: allocationMap

set $\boldsymbol{\mathcal{S}}$ = Sort GPPs in $\boldsymbol{\mathcal{M}}$ in decreasing order of CPU
for each user $u \in \boldsymbol{\mathcal{U}}$ with workload to be allocated **do**
    Starting with first GPP in $\boldsymbol{\mathcal{S}}$, allocate user workload W_u
    to GPP that will ~~accomodate~~ *accommodate* W_u, such that all GPPs are
    fully utilized.
    if If no GPP found **then**
        Activate a new GPP
        Allocate W_u to that GPP
    end if
end for

consolidation framework and compared with the traditional LTE system.

(i) Power Consumption: This is the total power consumed in the entire network measured in watts that considers the power consumption both from the BS cloud and radio side. Power consumption metric is important in showing the network that will have less OPEX and also ~~enviromentally~~ *environmentally* friendly with minimal CO_2 ~~emmissions~~ *emission*. Equation (3) is used to compute the total power consumption of the C-RAN.

(ii) Energy Efficiency (EE): The EE is defined as the ratio of average network throughput to power consumption in the network and is measured in bits/joule. The higher the EE the better the performance of the network. It is assumed that there are N available channels in every cell for transmission with each having bandwidth $BW = B/N$ where B is the cell bandwidth. In this regard, a channel means one PRB which is allocated to each user per scheduling interval. For simplicity it is assumed that different frequency bands are used by adjacent BS so inter channel interference (ICI) has been taken care of. Thus, the throughput of a user u can be formulated as [28]:

$$r_u = BW.log_2\left(1 + \frac{\eta_0.P_u}{d^\alpha}\right) \tag{13}$$

where α is the path-loss exponent and $\eta_0 = G_0/N_0$ includes the effect of antenna gain G_0 and thermal noise N_0, and d is the distance from the RRH to the user. P_u is the transmission power per user.

In addition, the overall cell/RRH throughput r_n can be written as:

$$r_n = B.\sum_{u \in \boldsymbol{\mathcal{U}}}^{N} log_2\left(1 + \frac{\eta_0.P_u}{d^{\alpha}}\right) \tag{14}$$

Based on (12) and (3), EE of a the total network in bits per joule is:

$$\eta_{EE} = \frac{\sum_{n=1}^{R} r_n}{P_{C-RAN}} \tag{15}$$

The aim is to improve η_{EE}, by reducing total power consumption through the proposed workload consolidation.

(iii) Resource Utilization: It is the ratio of the processing workload from cells to the maximum capacity of servers utilized in the network. It shows how efficiently resources are being utilized and this can be measured by using (11).

(iv) Statistical Multiplexing Gain, θ: The ratio of infrastructure (BBU servers) used in traditional LTE system to the infrastructure used in C-RAN. The higher the value the better the gain. Thus,

$$\theta = \frac{\#\text{servers in traditional LTE system}}{\#\text{servers in C-RAN}} \tag{16}$$

(v) Number of BBU Servers This is the total number of BBU servers used in the entire network according to traffic load. The C-RAN is expected to use less BBUs due to the workload consolidation mechanism. The number of servers in C-RAN are the output of running the workload consolidation algorithm.

4. SIMULATION RESULTS AND DISCUSSION

4.1. Simulation Settings

To analyse the performance of the proposed workload consolidation scheme for C-RAN, a simulation layout of 10 cells comprising of a maximum of 10 BBUs was considered. Bandwidth of 10 MHz was considered and up to 50 users in total are randomly generated per cell. The number of users in the cell follows the traffic profile in Fig. 1. Each user is allocated one PRB per transmission time interval (TTI), which is 1ms, in a proportional-fairness

manner. The GPPs for the BS cloud consists of ISS (Industry Standard Server) blade servers called Intel Xeon Processor E5540 [26][29] with Quard Core (45 GOPS per CPU) at its maximum efficiency (CPU frequency always at maximum). The total processing power for the server is 180 GOPS. Other parameters are shown in Table 1 from [3][24][28][29][30]. All results using the C-RAN workload consolidation are compared with the baseline LTE system which comprises of 10 individual BBU processing servers for 10 cells.

4.2. Results Evaluation

Fig. 6 shows the number of servers used in both cases versus average cell processing workload. The results show that as the average cell workload increases, more BBU servers are required in the C-RAN case. It can be seen that the C-RAN workload consolidation outperforms compared to the baseline in minimizing the number of servers used *as more GPPs are turned off during low traffic periods to save energy*. For example, the C-RAN workload consolidation scheme uses 8 servers for the 10 cells during peak load condition and 1 server for 10 cells during low traffic condition. *The baseline scheme has constant number of GPPs which is always 10.* On average, the proposed scheme uses 5 servers while the baseline system always requires the number of BBU servers equal to the number of cells which, in this case, is always 10 servers. Fig. 7 shows that on average the proposed scheme uses only 5 servers compared to baseline that uses 10 while during low ~~trafic~~ *traffic* and peak traffic conditions, the proposed scheme uses 10% and 80% of the servers respectively compared to the baseline scheme. *More servers are required during peak taffic because more user requests are processed in the BS cloud*

Table 1: Parameters used in the simulations.

Parameter	Value
GPP model	Xeon Processor E5540
GPP GOPS (C_{cap})	180 GOPS
GPP lower threshold	30 % of 180 = 54 GOPS
GPP upper threshold	90 % of 180 = 162 GOPS
GPP idle power $P_0^{GPP_i}$	120 Watts
GPP maximum power $P_{max}^{GPP_i}$	215 Watts
Server power gradient $\triangle_P^{GPP_i}$	0.44
BS idle power P_0	324 Watts
BS gradient slope $\triangle_P$	4.2
Cloud cooling power $P_{cooling}$	500W
P_{base} *[31]*	*118.330W*
P_{config} *[31]*	*5.29W*
BS maximum output power P_{max}	46 dBm
Bandwidth B	10 MHz
No. of antennas Ant	2
Modulation K	4 bits (16-QAM)
Coding rate D	1
MIMO layers L	2
Number of users per cell	up to 50
Number of cells, R	10
Cell radius	500m
Inter-BS distance	>1 km
BS antenna gain G_0	16 dBi
Noise Power N_0	-141 dBm/Hz
Pathloss Exponent, α	4

Fig. 8 shows the statistical multiplexing gain and it is higher than one meaning that the proposed scheme uses less number of servers than the baseline approach. *This means more operational expenditure (OPEX) and capital expenditure (CAPEX) can be reduced and also the electricity bill as less BBUs are being utilized using C-RAN* During low traffic loads, the gain is higher, (eg. 10), and the ratio of baseline servers to C-RAN servers used for the same traffic is 10:1, i.e., the number of servers used in baseline is 10 times the num-

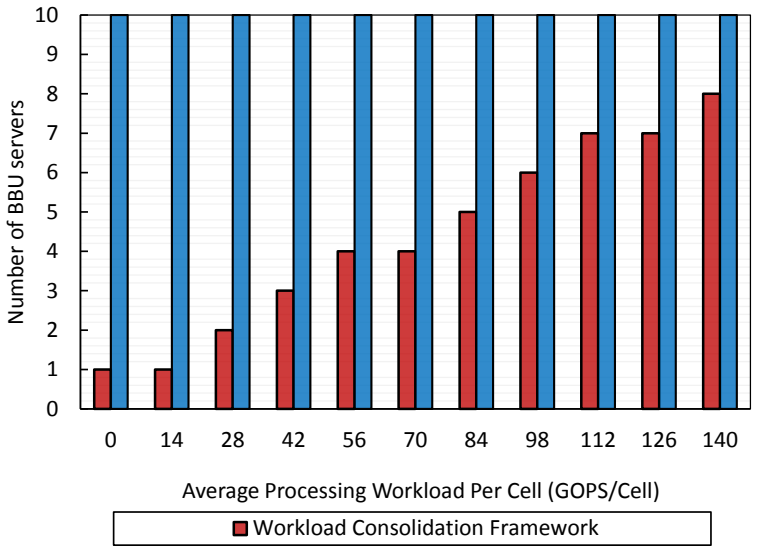

Figure 6: Number of BBU servers used versus processing workload.

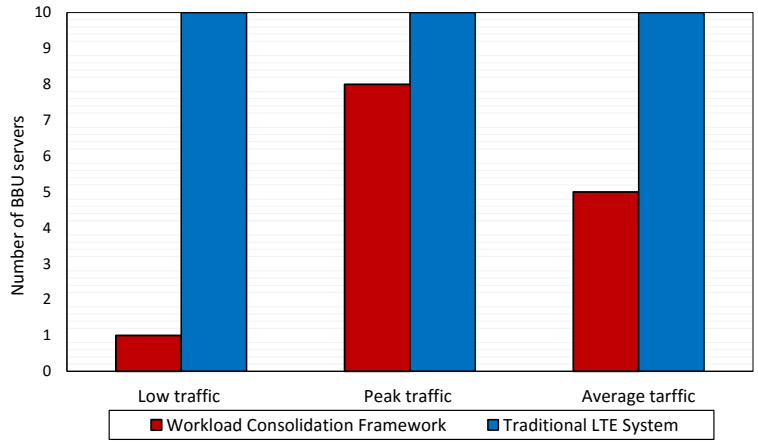

Figure 7: Daily average number of servers for different traffic load conditions.

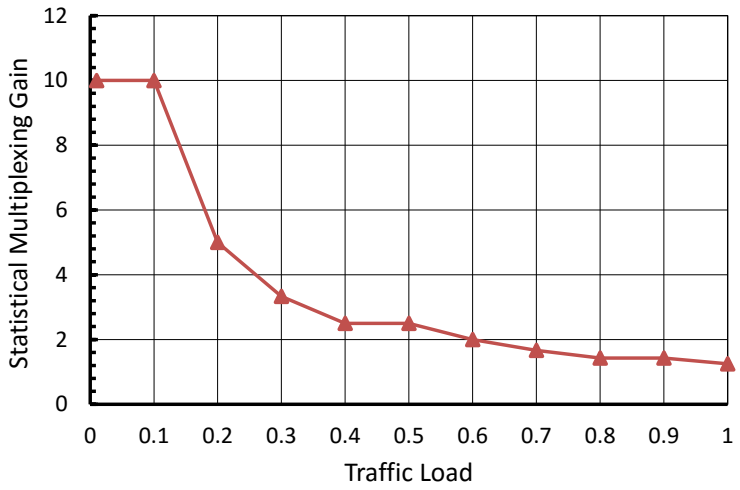

Figure 8: Statistical multiplexing gain of C-RAN.

ber of servers used in C-RAN. As the traffic load increases, the multiplexing gain decreases because more servers are gradually activated in C-RAN. At 100% traffic load, the gain is 1, which means the ratio of baseline servers to C-RAN servers used for the same traffic is 1 is to 1. *This is proof to show that C-RAN can save much on the number of GPPs used as compared to the baseline scheme*

Fig. 9 shows the power consumption for both cases versus traffic load. All or part of the baseband can be processed in the cloud for C-RAN. *When all baseband is processed in the BS cloud, it is called full centralization and when part of the baseband is processed in the cloud and some baseband processed in the RRH, it is called partial centralization.* For C-RAN, a certain percentage of baseband processing was moved to the cloud to see the effect on overall power consumption. The advantage of leaving some baseband tasks in the radio side on RRHs is to reduce the high bandwidth baseband signals transported between the BS cloud and the radio side that can cause high cost in fiber. As shown in the graph, as more baseband processing is moved to the cloud, more power savings are gained since more workload is consolidated and shared between servers. *In this case, the RRH will consume a lot of power as it process more baseband.* The power consumption for both systems increases with the increase in traffic load. During the peak load, maximum power was consumed for all systems because more BBU servers are utilized and consume more power due to load dependence of traffic and

power consumption. The baseline system consumes more power as expected since all BBUs are always on.

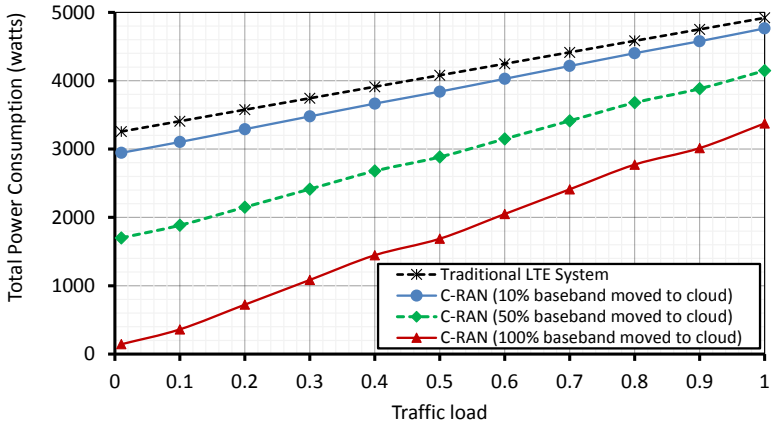

Figure 9: Power consumption versus traffic load.

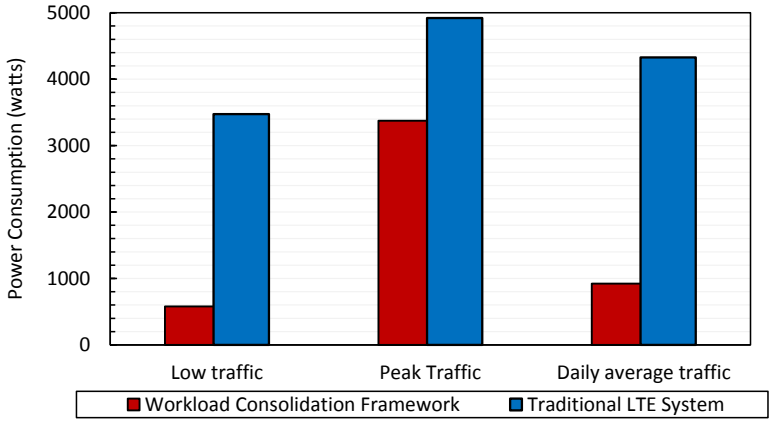

Figure 10: Daily average, peak and low traffic power consumption.

Fig. 10 shows the power consumption on daily basis. During low traffic, the proposed scheme saves up to 80% of energy compared to baseline system since at low traffic more BS traffic can be processed by fewer servers. However, during peak time, the proposed scheme saves 12% of energy compared to baseline system. On daily average, the workload consolidation model saves 38% of power compared to baseline system. *This shows more savings are*

achieved with the proposed scheme with reduced electricity bill especially in low traffic periods since BSs are hardly operating at peak traffic in real life.

Fig. 11 shows EE performance for increasing cell load for baseline and C-RAN workload consolidation scheme with 100% baseband moved to the cloud.*The higher the EE, the better the performance of the system.* In the latter, the EE improved because less power is consumed through workload consolidation as traffic from various cells are aggregated in fewer number of servers *where the numerator of the EE equation becomes smaller due to low energy consumption*. This contrast with the baseline system which has approximately constant and lower EE *at 0.05* due to increased power consumption with all ten BBUs always on. At low traffic load, C-RAN EE outperforms baseline by a much larger factor of four since at low traffic a single server can process traffic from many cells and other servers are turned off hence reducing energy consumption. As traffic increases gradually, EE for C-RAN drops gradually to almost a constant since more servers are turned on and utilised to process increased cell traffic hence consuming more energy.

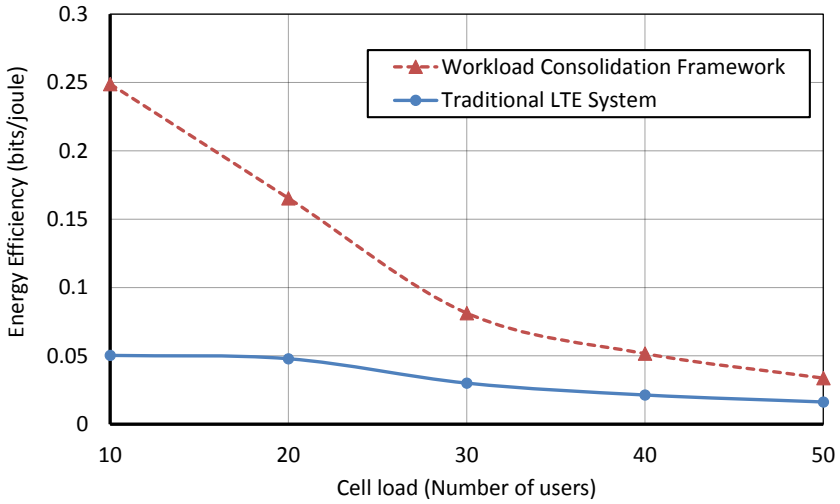

Figure 11: Energy efficiency versus cell load.

Fig. 12 shows the resource utilization in the network versus cell load (average number of users within the cell). It can be clearly seen from the diagram that for both systems resource utilization increases as the traffic load increase because servers process more traffic. The figure shows that for the same cell load, the proposed scheme has a higher utilization than the

baseline system. This is because fewer servers are processing all traffic from all cells in C-RAN whereas for the baseline, servers are not shared regardless of traffic load. At peak traffic, the utilization for both systems is 100% as all servers are utilized. *C-RAN has high utilization because the BBU servers are shared in the BS cloud. A single BBU server can process traffic requests from multiple RRHs at the same time.*

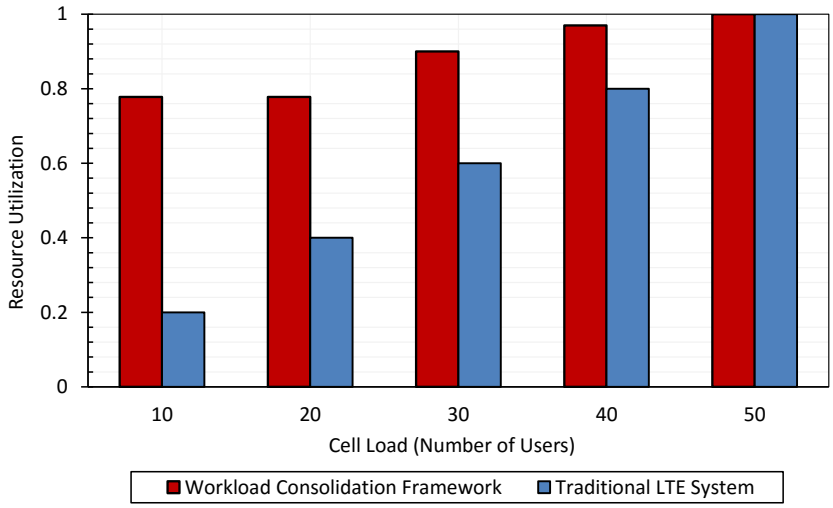

Figure 12: Resource utilization versus cell load.

5. CONCLUSION

This paper proposed a workload consolidation ~~technique~~ framework model for minimizing energy consumption in cloud radio access network (C-RAN) by reducing the number of baseband processing servers used. The number of computing servers are reduced by matching the right amount of baseband processing with traffic load with servers running at peak utilization. Idle servers can then be switched off to save energy. Extensive simulation and experimental results demonstrate that the proposed C-RAN workload consolidation scheme achieves an enhanced energy performance compared to the traditional LTE system. The proposed workload consolidation framework can save up to 80% of energy compared to LTE system. In future, the C-RAN workload consolidation scheme will be extended to heterogeneous networks

(HetNets) to include switching off the radio front end of small cells in the radio side in relation to traffic to further save more energy and improve EE.

References

[1] Huawei, "5G: a technology vision," in *White Paper of Huawei Tech'13*, 2013.

[2] I. Chih-Lin, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: A 5G perspective," *Comm. Mag., IEEE*, vol. 52, pp. 66–73, 2014.

[3] China Mobile, "C-RAN road towards green radio access network," in *C-RAN International Workshop*, 2010.

[4] T. Sigwele, P. Pillai, and Y. F. Hu, "iTREE: Intelligent Traffic and Resource Elastic Energy Scheme for Cloud-RAN," in *FiCloud/IEEE*, 2015, pp. 282–288.

[5] T. Sigwele, P. Pillai, and Y. F. Hu, "Call admission control in cloud radio access networks," in *FiCloud/IEEE*, 2014, pp. 31–36.

[6] G. Auer, V. Giannini, C. Desset, "How much energy is needed to run a wireless network?," *Wireless Comm. Mag., IEEE*, vol. 18, no. 5, pp. 40–49, 2011.

[7] E. Yaacoub, "A practical approach for base station on/off switching in green lte-a hetnets," in *Proc. 10th WiMob/IEEE*, 2014, pp. 159–164.

[8] J. F. Cheng, H. Koorapaty, P. Frenger, D. Larsson, and S. Falahati, "Energy efficiency performance of lte dynamic base station downlink dtx operation," in *Proc. 79th VTC Spring/IEEE*,2014, pp. 1–5.

[9] L. Suarez, L. Nuaymi, J. Bonnin, "Energy-efficient BS switching-off and cell topology management for macro/femto environments," *Computer Networks*, vol. 78, pp. 182–201, 2015.

[10] C. Murthy and C. Kavitha, "A survey of green base stations in cellular networks," *International Journal of Computer Networks and Wireless Communications*, vol. 2,no. 2, pp. 232–236, 2012.

[11] T. Sigwele, A. S. Alam, P. Pillai, and Y. F. Hu, "Evaluating Energy Effecient Cloud Radio Access Networks for 5G," in *GREENCOM*, Dec. 2015.

[12] S. Namba, T. Warabino, and S. Kaneko, "Bbu-rrh switching schemes for centralized ran," in *CHINACOM/IEEE*, 2012, pp. 762–766.

[13] Z. Kong, J. Gong, C. Z. Xu, K. Wang, and J. Rao, "ebase: A baseband unit cluster testbed to improve energy-efficiency for cloud radio access network," in *ICC/IEEE*, 2013, pp. 4222–4227.

[14] T. Zhao, J. Wu, S. Zhou, and Z. Niu, "Energy-delay tradeoffs of virtual base stations with a computational-resource-aware energy consumption model," in *ICCS/IEEE*, 2014, pp. 26–30.

[15] C. Liu, K. Sundaresan, M. Jiang, S. Rangarajan, and G. K. Chang, "The case for re-configurable backhaul in cloud-ran based small cell networks," in *INFOCOM/IEEE*, 2013, pp. 1124–1132.

[16] S. Namba, T. Matsunaka, T. Warabino, S. Kaneko, and Y. Kishi, "Colony-ran architecture for future cellular network," in *FutureNetw/IEEE*, 2012, pp. 1–8.

[17] C. Liming, H. Jin, H. Li, J. Seo, Q. Guo, and V. Leung, "An energy efficient implementation of C-RAN in HetNet." in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, pp. 1-5. IEEE, 2014.

[18] H. Nguyen, and L. Bao Le, "Cooperative transmission in cloud RAN considering fronthaul capacity and cloud processing constraints." in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1862-1867. IEEE, 2014.

[19] M. Khan, R. S. Alhumaima, and H. S. Al-Raweshidy, "Reducing energy consumption by dynamic resource allocation in C-RAN." in *Networks and Communications (EuCNC)*, 2015 European Conference on, pp. 169-174. IEEE, 2015.

[20] Z. Kong, J. Gong, C. Xu, K. Wang, and J. Rao, "ebase: A baseband unit cluster testbed to improve energy-efficiency for cloud radio access network." in *2013 IEEE International Conference on Communications (ICC)*, pp. 4222-4227. IEEE, 2013.

[21] C. Lui, K. Sundaresan, M. Jiang, S. Rangarajan, and G. K. Chang, "The case for re-configurable backhaul in cloud-RAN based small cell networks." in *INFOCOM, 2013 Proceedings IEEE*, pp. 1124-1132. IEEE, 2013.

[22] P. Li, T. Chang, and K. Feng, "Energy-efficient power allocation for distributed large-scale MIMO cloud radio access networks." in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1856-1861. IEEE, 2014.

[23] M. Yingna, M. Peng, Z. Zhao, and Z. Zhou, "Optimization of Simultaneous Wireless Information and Power Transfer in Cloud Radio Access Networks." in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, pp. 1-5. IEEE, 2016.

[24] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, H. Holtkamp, W. Wajda, D. Sabella, F. Richter, M. J. Gonzalez, "Flexible power modeling of lte base stations," in *WCNC/IEEE*, 2012, pp. 2858–2862.

[25] Extron, "Fiber Power Budget," *Online,[Visited: April 2016]*, Available: `http://www.extron.com/product/fibercalculator.aspx`.

[26] FIT4Green, "Presentation of full-featured federated energy consumption models ," *Online, [Visited: May 2016]*, Available: `http://www.fit4green.eu/sites/default/files/attachments/documents/D3.3_final.pdf`.

[27] T. Werthmann, H. Grob-Lipski, and P. Proebster, "Multiplexing gains achieved in pools of baseband computation units in 4G cellular networks," in *Proc. 24th PIMRC/IEEE*, Sep. 2013, pp. 3328?-3333.

[28] A. S. Alam, L. S. Dooley, and A. S. Poulton, "Energy efficient relay-assisted cellular network model using base station Switching," in *Proc. 24th GLOBECOM Workshops/IEEE*, 2012, pp. 1155–1160.

[29] Intel Corporation, "Intel Xeon Processor 5500 series ," *[Online]*, *[Visited: May 2016]*, Available: `http://download.intel.com/support/processors/xeon/sb/xeon_5500.pdf`.

[30] H. Holtkamp, G. Auer, V. Giannini and H. Haas, "A Parameterized Base Station Power Model," in *IEEE Comm. letters*, vol. 17, no. 11, pp. 2033–2035, 2013.

[31] K. Fabian, S. Melnikowitsch, and D. Hausheer, "Measuring and modeling the power consumption of OpenFlow switches," in Proc. 10th International Conference on Network and Service Management (CNSM) and Workshop/IEEE*, 2014.*